Analysis of SITA queues with many servers and space-time geometry

Eitan Bachmat [*] Assaf Natanzon [†]

1. INTRODUCTION AND PRELIMINARIES

SITA queues were introduced in [4] as a means for reducing job size variance at individual hosts in a server farm. It turns out that SITA queues are mathematically very interesting. For example, they satisfy a duality that is typical of automorphic forms in number theory. This leads to useful queueing theoretic insights, [1] and is also related to interesting number theoretic questions, [7], [5].

In this paper we will consider other aspects of SITA queues which turn out to be related to two dimensional Lorentzian geometry. In particular we will be interested in the behavior of SITA queues as $h \to \infty$. The tail of the waiting time function was studied in detail in [8]. We will concentrate on the average waiting time $E(W)$, rather than the tail since it leads to some interesting analogy and insights.

A SITA queue consists of h hosts, numbered $1, ..., h$ and a set of cutoffs $0 = s_0 < s_1 < s_2 < ... < s_{h-1} < s_h = \infty$. Assuming that all hosts are identical, and given a job of size x, we assign it to host i such that $s_{i-1} \le x < s_i$.

More generally, if the hosts are not identical, we assume that each job is assigned some size x and we proceed as before. We will assume that the hosts are *coupled* which means that the time it takes a job of size x to process at host j is a function $t_j(x_i)$. If the hosts are *identical* we take t_i to be the identity function. We say that host strengths are *linearly coupled* if there exist constants c_j, $j > 1$ such that $t_j(x)/t_1(x) = c_j$ for all x. In this case, we again take t_1 to be the identity (job sizes are measured by the time they take on host 1).

2. CHOOSING GOOD CUTOFFS

One of the major problems in SITA queues is finding the optimal cutoffs, given all the other parameters, i.e., X, t_j, λ and the target function $E(W)$. Given a SITA system, let $s_1^{qbal}, \dots, s_{h-1}^{qbal}$ be the cutoffs which balance average queue length at the hosts. In some cases it may not be possible to find cutoffs which completely balance (equalize) average queue length. In those cases, we let s_i^{qbal} be those cutoffs which minimize the maximal average queue length among all hosts. We have the following simple and basic result.

THEOREM 1. *Consider a* SITA *queue which satisfies the assumptions of Little's law. For any,* $s_1, \dots, s_{h-1}$ *we have*

$$hE(W)(f, h, \rho, s_1, \dots, s_{h-1}) \ge E(W)(f, h, \rho, s_1^{qbal}, \dots, s_{h-1}^{qbal}) \quad (1)$$

Proof: Given any set of cutoffs $s_1, \dots, s_{h-1}$, let p_i denote the portion of jobs which arrive at host i of the SITA system, namely, the probability that $s_{i-1} \le s < s_i$. We have $E(W) = \sum_i p_i E(W_i)$ where $E(W_i)$ is the average waiting time at the i'th host. For the same set of cutoffs we may also define $M(s_1, \dots, s_{h-1}) = Max_i \; p_i E(W_i)$. Obviously,we have for any set of cutoffs $M(s_1, \dots, s_{h-1}) \le E(W)(s_1, \dots, s_{h-1}) \le hM(s_1, \dots, s_{h-1})$. We note that the arrival rate to host i is λp_i, where λ is the system arrival rate. Let $E(Q)_i$ be the average queue length at host i. By Little's law, we have that

$$E(Q)_i = \lambda p_i E(W)_i \quad (2)$$

By definition $s_1^{qbal}, \dots, s_{h-1}^{qbal}$ minimizes the function

$$Max_i \; E(Q_i)$$

which by (2) is the same as minimizing M.

We have

$$E(W)(s_1, \dots, s_{h-1}) \ge M(s_1, \dots, s_{h-1}) \ge$$

$$M(s_1^{qbal}, \dots, s_{h-1}^{qbal}) \ge E(W)(s_1^{qbal}, \dots, s_{h-1}^{qbal})/h$$

Multiplying both ends of the equation by h we get the desired result. *q.e.d.*

3. THE MANY SERVER CASE

The result above shows that cutoffs which balance queue length are never far from optimal when minimizing average waiting time, though, the average queue lengths for the optimal cutoffs can be highly unbalanced. This occurs for bounded Pareto distributions P_{α_n, l_n} with $\alpha_n \to 2$ from below and $l_n \to \infty$ at a fast enough pace. The same sequence shows that the constant h is in general, the best possible. However, as we will show, for a fixed bounded distribution the queue length balancing policy becomes optimal.

From now on, we will assume Poisson arrivals, that the job sizes are drawn from an i.i.d. generic distribution X, and (at first) a fixed rate λ of arrivals regardless of the number of hosts. Our target function will be the *average normalized waiting time* $E(N) = E(W)/E(X)$. We could have used

[*] Eitan Bachmat, Department of Computer science, Ben-Gurion University, Beer-Sheva, Israel, 84105. ebachmat@cs.bgu.ac.il

[†] Assaf Natanzon, EMC Corp., Ramat Gan, Israel and department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel, 84105. assaf.natanzon@emc.com

$E(W)$ just as well, however, is invariant under change of time units and also the duality of SITA queues, hence it is preferable.

We consider the case where X_p is some fixed distribution of job sizes in the range $[1, p]$ with a smooth strictly positive density function f_p, a fixed arrival rate and the number of hosts h tending to infinity (becoming large). This asymptotic regime, was first considered in [3] under the name, the server expansion metric. Let $[a, b]$, be some fixed interval. Let $g : [a, b] \longrightarrow [1, p]$ be an increasing function with $g(a) = 1$ and $g(b) = p$. We will think of such functions g as providing a formula for choosing the cutoffs. We say that $1 < s_{1,h} < \ldots < s_{h-1,h} < p$ is a *compatible family of cutoffs*, if there is a function g as above such that

$$s_{i,h} = g(a + (b - a)i/h) \tag{3}$$

Let $E(N)(g, X_p, h, \rho)$ be the mean normalized waiting time for a SITA system with generic job size distribution X_p, arrival rate λ, h hosts and cutoffs given by (3). We say that g is an *h-asymptotically optimal cutoff formula* if for any other family of cutoffs $s_{i,h}$, compatible or not, for any λ and for any $\varepsilon > 0$ we have

$$E(N)(g, X_p, h, \rho)/E(N)(s_{i,h}, X_p, h, \rho) < 1 + \varepsilon$$

for h large enough. The following theorem characterizes the h-asymptotic performance of a SITA queue.

Theorem 2. *Consider a SITA queue, with h identical hosts, a fixed arrival rate λ and with a job size distribution given by an infinitely differentiable density function f on the interval $[1, p]$. There exists a unique, differentiable, strictly increasing function g which is an h-asymptotically optimal cutoff formula. The function g is a solution to the differential equation*

$$f(g(x))g(x)g'(x) = c \tag{4}$$

where $c > 0$ is some constant, and g satisfies the boundary conditions $g(a) = 1$ and $g(b) = p$. The equation says that in the h-asymptotic regime the asymptotically optimal cutoffs, balance the queue length in each server and also balance load. The same conclusions hold for the case in which the arrival rate is$\lambda = \tilde{\lambda}h$, i.e., grows linearly with the number of hosts.

We have

$$E(N)^{opt} \sim \frac{\rho}{2h(1 - \rho/h)}(b - a)^2 c^2 \frac{1}{E(X)^2} \tag{5}$$

where $\rho = \lambda E(X)$.

Proof: Each host is responsible for job sizes in a small range given by equation (3). Let $dx_i = (b - a)/h$. Let $ds_i = s_i - s_{i-1}$. We have

$$ds_i = s_i - s_{i-1} = g(a + (b - a)i/h) - g(a + (b - a)(i - 1)/h)$$

$$\sim g'(x_i)(b - a)/h = g'(x_i)dx_i$$

where $f(h) \sim g(h)$ denotes the assertion that $f(h)/g(h) \to 1$ as $h \to \infty$. Let k be a function on the positive reals and I an interval in the positive reals. Consider the incomplete Mellin transform of f w.r.t. I, given by $L_{f,I}(s) = \int_I t^{s-1} f(t)dt$. Let $I_i = [s_{i-1}, s_i)$ denote the interval of job sizes assigned to host i. If $I = (0, \infty)$, we obtain the *Mellin transform* $L_f(s)$.

Since f is assumed to be bounded from above, we see that the utilization ρ_i on each host tends to zero and hence $1 - \rho_i \to 1$. From the Pollaczek-Khinchine formula we verify that the contribution of host i to the computation of $E(N)$ is given by

$$E(N_i) \sim \frac{\rho}{2} \frac{L_{f,I_i}(1) L_{f,I_i}(3)}{(L_f(2))^2}$$

We have

$$L_{f,I_i}(3) \sim g'(x_i) f(g(x_i)) g(x_i)^2 dx_i$$

Similarly

$$L_{f,I_i}(1) \sim g'(x_i) f(g(x_i)) dx_i$$

Summing over i and recalling that $dx_i = \frac{b-a}{h}$ we have the Riemann sum expression

$$E(N) = \sum_i E(N_i)$$

$$\sim \frac{\rho}{2} \frac{\sum_i (f(g(x_i))g(x_i))^2 g'(x_i)^2 dx_i^2}{(L_f(2))^2}$$

$$= \frac{\rho(b - a)}{2h(L_f(2))^2} \sum_i (f(g(x_i))g(x_i))^2 g'(x_i)^2 dx_i$$

Taking the limit as h goes to infinity we see that

$$E(N) \sim \frac{\rho(b - a)}{2h(L_f(2))^2} \int_a^b (f(g(x))g(x))^2 g'(x)^2 dx$$

We need to minimize the integral expression. Given h hosts and optimal cutoffs $s_1, \ldots, s_{h-1}$, we define as described before, the non decreasing curve g_h by extrapolating linearly between the cutoff values. Since the set of non decreasing curves is compact in the C_0 topology, a subsequence of the curves g_h converges in that topology to a curve g_∞.

We note that the functional of interest has the form

$$F(x, g, g') = \tilde{L}(g)g'^2$$

which is a special case of the more general form

$$F(x, g, g') = \tilde{L}(x, g)g'^2$$

This bares a strong resemblance to functionals of the form

$$F(x, g, g') = \tilde{L}(x, g)\sqrt{g'}$$

which is the proper time of a causal curve in a two dimensional space-time with a Lorentzian metric. The minimizers of the proper time functional are the geodesics. The main difference between these functionals is that the function z^2 which is applied to g' in the first case is strictly convex, while the function $\sqrt{z}$ which is applied to g' in the second case is strictly concave. This difference in particular is responsible for the fact that a generic solution to the Euler-Lagrange equation in the first case is a minimum while in the second case it is a maximum.

using the same arguments as in [6] for the proper time functional, we conclude that our functional is lower semicontinuous in the C_0 topology, hence g_∞ is a minimizer and consequently an h-asymptotically optimal cutoff formula exists. The argument that the limiting curve g_∞ is the graph of a differentiable strictly increasing function comes as in the case of geodesics from the theory of existence and uniqueness of ODE, see again [6] or [2] for similar arguments. Since x does not figure explicitly in

$$F = \tilde{L}(g)g'^2 = f^2(g)g^2 g'^2$$

the Euler-Lagrange equation which a minimizing function must satisfy degenerates to the Beltrami equation

$$g'\frac{\partial L}{\partial g'} - L = c_1$$

where c_1 is some constant, or in this particular case

$$2g'^2\tilde{L} - \tilde{L}g'^2 = \tilde{L}g'^2 = F(g, g') = c_1$$

The constant c_1 must be positive because L and g' are. The equation (4) is obtained by taking the square root. Since as h becomes large $E_i(N) \sim F(g, g')(a + i(b-a)/h)$ we see that we may interpret the Beltrami equation as load balancing the contributions of the different hosts to the average waiting time. By Little's law this also means balancing the average queue lengths at the hosts. The equation (4) itself describes the relative load at each host, hence this is also load balanced.

Plugging the equation into the computation of $E(N)$ and recalling that $\rho/h \to 0$ when ρ is fixed, gives the expression in (5).

The solution is unique since there is a unique load balancing solution for each h. We can also argue that it is unique since it also solves the geodesic equation in Minkowski space.

If we consider the case where the utilization per host remains constant we obtain a slightly more complicated functional

$$\frac{\rho}{2}(b-a)\frac{1}{(m_1^f)^2}\int_a^b \frac{(f(g(x))g(x))^2 g'(x)^2}{1-\rho(f(g(x))g(x))}dx$$

but, the solution to the corresponding Beltrami equation is the same as before. *q.e.d.*

We may consider linearly coupled servers in the context of the theorem above. We will need to assume that the speed up constants $c_{i,h}$, come from a smooth function $c(x)$ by setting $c_{i,h} = c(a + i(b-a)/h)$. This amounts to a restriction that hosts which handle nearby job size intervals have nearly equal processing speed. Under this assumption the corresponding functional for determining asymptotically optimal cutoffs will have the general form

$$A(x)B(g)(g')^2$$

. Since we have a product form $A(x)B(g)$, we can make separate monotone changes of variables to x and g (or y in the usual coordinate notation), which reduces the functional to the identical server case. We get after the change of coordinates the previous results about uniqueness of the solution and having a conserved quantity (A Beltrami equation). We note that the same holds in Lorentzian geometry, the equation $A(x)B(g)\sqrt{g'}$ is the geodesic equation of flat Minkowski space, but in non-standard coordinates where one applies monotone changes of variables to the x and y coordinates separately. Essentially we are still in the Minkowski space case.

We may consider, more generally, the case of (non-linearly) coupled servers. To consider the large h limit, we again assume that the functions $t_{i,h}(t)$, lead in the limit to a continuous function $T(x, g)$. The resulting functional will have the form

$$\tilde{L}(x, g)(g')^2$$

This is similar to the proper time functional in Lorentzian geometry, however, this time the Lorentzian metric is not Minkowski, it is in general a *curved* space-time, which no coordinate change will convert to Minkowski space. In analogy with Lorentzian geometry, we do not expect in this case to always have a unique solution and examples can be constructed where the solution is not unique. The solution need not be unique since the Euler-Lagrange equation is not a Beltrami type equation anymore and we do not have a balanced quantity. Going from the linearly coupled model to the coupled model is analogous to introducing curvature in space-time!

Acknowledgements: Eitan Bachmat has been supported by ISF grant number 580/11.

4. REFERENCES

[1] Bachmat E. and Sarfati H. 2010. Analysis of Size Interval Task Assignment SITA policies. *Performance Evaluation*, 67(2), 102-120, 2010.

[2] J.D. Deuschel and O. Zeitouni, Limiting curves for iid records, *Annals of probability*, 23, 852-878, 1995.

[3] M. Harchol-Balter, Task assignment with unknown duration, *Journal of the ACM*, vol. 49(2), 260-288, 2002.

[4] M. Harchol-Balter, M. Crovella and C. Murta, On choosing a task assignment policy for a distributed server system, *IEEE Journal of parallel and distributed computing*, Vol. 59, 204-228, 1999.

[5] J. Jung, On sparsity of positive definite automorphic forms within a family, arXiv:1201.0429v1, 2012.

[6] R. Penrose, Techniques of differential topology in relativity, Regional conference series in applied mathematics Vol.7, SIAM, 1972.

[7] P. Sarnak, Letter to E. Bachmat on positive definite L-functions, available at http://publications.ias.edu/sarnak/paper/511

[8] B. Zhang and B. Zwart, Steady-state analysis for multi-server queues under size-interval task assignment in the Quality-Driven regime, technical report.